NTNU
Norwegian University of
Science and Technology

Speech adaptation of special voice classes

Bjørnar Grip Fjær

Master of Science in Electronics
Submission date: June 2011
Supervisor: Torbjørn Svendsen, IET

Norwegian University of Science and Technology
Department of Electronics and Telecommunications

Problem Description

Most speech recognizers are trained on databases with speech from young to middle age adult speakers. There are, however, many groups of people who have voices that fall outside this category. Kids voices will, for example, differ from adult voices in that the base frequency will be higher than for adults, and the formant placements will be different. For older speakers, there are also systemic differences in the speech. Since the short time frequency analysis uses estimates of the spectral envelope, changes in the formant placements, for example due to the shorter vocal tract in children compared to adults, will be an important reason for the difference. The effect of such systematic differences can be simulated, so that a training database can be created for different voice classes that are not covered by available databases. This will be a substantially more effective (and cheap) alternative to building a new database for these voice classes. The assignment is to find what factors characterise speech from different voice classes, build a simulated database based on a standard database and a model of these factors, and evaluate the performance of a speech recognizer trained on these simulated databases.

Assignment given: 13. January 2011
Supervisor: Professor Torbjørn Svendsen

Abstract

Most automatic speech recognition systems are based on statistical models that require training. While these types of systems have reached recognition rates that are sufficient for many purposes, they perform poorly for speaker types that are not present in the training material. Children are often absent from training material for speech recognizers, and creating good training material for children can be difficult and expensive.

To address this issue, this thesis focuses on using adult training material to train a recognizer for children by adapting the training material during training. Instead of performing speaker-dependent adaptation during recognition, where computational power may be scarce, and responsiveness may be essential, adaptation is performed during training towards a class of speakers.

Using a combination of vocal tract length normalization (VTLN) and cepstral mean normalization during training, promising results have been obtained. In a connected-digits task, a reduction in errors as high as 70% was shown, with a reduction of almost 50% in a large vocabulary task. Using VTLN to warp the same training material several times, combining these warped materials to train one recognizer, a similar reduction in errors was shown, but with an increased robustness indicating a less speaker-dependent system. It is also shown that a piecewise linear warping method is better suited to warp adult speech to child speech, than a bilinear warping method.

Preface

The work behind this report was done in the spring semester of 2011 at the Norwegian University of Science and Technology, Institute of Electronics and Telecommunications. This report is the final work of my master thesis in electronics engineering.

I would like to thank my supervisor Professor Torbjørn Svendsen for his valuable feedback and interesting discussions.

Contents

List of Figures

List of Tables

Glossary

ASR Automatic Speech Recognition. 1, 5, 6, 11, 19, 36

CMN Cepestral Mean Normalisation. 3, 19, 28, 32, 36, 37, 51

CMU Carnegie Mellon University. 21, 22

DCT Discrete Cosine Transform. 10

DFT Digital Fourier Transform. 9

GMM Gaussian Mixture Model. 11, 12

HMM Hidden Markov Model. 11, 12, 26

HTK Hidden Markov Model Toolkit. 6, 10, 15, 19, 23, 29, 35

IIR Infinite Impulse Response. 13

MFCC Mel-frequency cepstral coefficient. 2, 9, 11, 12, 16, 25

VTLN Vocal tract length normalization. 2, 3, 13, 16, 25, 28, 32, 35–37

WER Word error rate. 7, 8, 25, 26, 28, 30–32, 35–37, 46–48, 51, 52

Chapter 1

Introduction

Todays speech recognition systems are overwhelmingly based on a statistical approach, where a database of recorded speech with known content is used to train an Automatic Speech Recognition (ASR) system and create models that are later used to recognize speech based on the same statistical principles. This approach has proven to be fairly successful in speech recognition, but has certain obvious weaknesses. A significant problem is that any type of speech not present in the training material will suffer from poor performance in such a recognizer, due to the statistical properties not being present in the training material.

This thesis describes efforts to increase the recognition rate for speakers with uncommon speech types, specifically children, who are often underrepresented in training databases. Studies have shown that the recognition rate for children suffer when they are not included in the training material [Wilpon and Jacobsen, 1996]. Creating good training material for children is a difficult and time consuming task. Depending on the age, the child may be unable to read, in which case the sentences would have to be prompted by an adult speaker, which could lead to the child imitating the adult and pronounce it differently that he or she normally would. If the child reads poorly, it could also affect the pronunciation. Children also tire more quickly than adults and tend to lose focus faster, so making training material for children often involve development of special interfaces to keep the children interested [Kazemzadeh et al., 2005, Shobaki et al., 2000].

To address the problem of creating training material for children, attempts are made to adapt adult training material to serve as child training material. The focus in this thesis is on acoustical adaptation of the speech, but it has been shown that there is a necessity for special language models for children as well [Das et al., 1998]. By adapting the material during training, the computational costs occur mainly during training, where computational

power is cheap.

The primary adaptation method used is Vocal tract length normalization (VTLN). This is a method used to adjust for spectral differences between speakers by warping of the frequency spectrum [Lee and Rose, 1996]. Studies show that the spectral differences between children and adults can be very significant [Lee et al., 1999, Narayanan and Potamianos, 2002], so adjusting for these differences is important when comparing adult and child speech. The age of the speaker can also be related to the length of the vocal tract [Fitch and Giedd, 1999].

Earlier work has shown that using linear VTLN on adult training material can significantly increase the performance of recognizers when tested on children [Elenius, 2010, Potamianos and Narayanan, 2003], with reduction in errors as high as 50% compared to unadapted adult training material. The use of phoneme-dependent adaptation showed varying results, but the lack of reduction in errors was primarily blamed on the huge search space and the problems of estimating good warp factors in a large search space. To solve this issue, adaptation based on phonetic classes was suggested as a suitable substitute, which would reduce the search space significantly.

In [Stemmer et al., 2003] using non-linear VTLN was attempted to further increase the recognition rate for children. While the non-linear approach attempted only showed a marginal increase in recognition rate, it did not give any conclusive answers, but suggested that a non-linear approach may outperform linear adaptation.

Many of the experiments performed have been on small vocabulary, connected digits task. In this thesis, the performance of adapting adult training material to work as child training material on a large vocabulary task is evaluated.

This thesis is the continuation of a specialization project done in the autumn semester of 2010, which focused on the use of VTLN to improve the recognition rates for children using adult training data [Fjær, 2010].

1.1 Report outline

The rest of this thesis is organized as follows. In Chapter 2 the basic recognition system is explained, and some information about the specific software used in this thesis.

An explanation of how the performance of the system is evaluated is given in Chapter 3.

Chapter 4 explains the process of creating Mel-frequency cepstral coefficient (MFCC)s, the basic features used to model the speech used in these

experiments.

In Chapter 5 the statistical techniques used to create a model to be used for recognition are explained.

VTLN is explained in chapter 6. Here, different methods of performing VTLN is detailed, as well as how VTLN is implemented in the recognition systems used in these experiments.

Chapter 7 explains the principles behind Cepestral Mean Normalisation (CMN).

The speech corpora used in these experiments are detailed in 8. An overview is given of the different sets used for training and recognition.

The experiments performed and the results obtained are shown in 9. Due to the experiments performed stemming from results of previous experiments, a small discussion of the results is usually given to keep the natural progression of experiments.

Chapter 10 discusses the results found and details some problems, and some further improvements that could be made.

In Chapter 11 the conclusions are given about what the results have shown and what purpose they might serve.

Chapter 2

Recognition system

Figure 2.1 shows the training process of an ASR system. Training material containing recordings of speech and a transcription of the content in each recording is needed. Feature extraction is performed on the speech, passing it to the modelling stage together with the transcribed text. The transcribed text is used to make a decision about which model each of the features belong to.

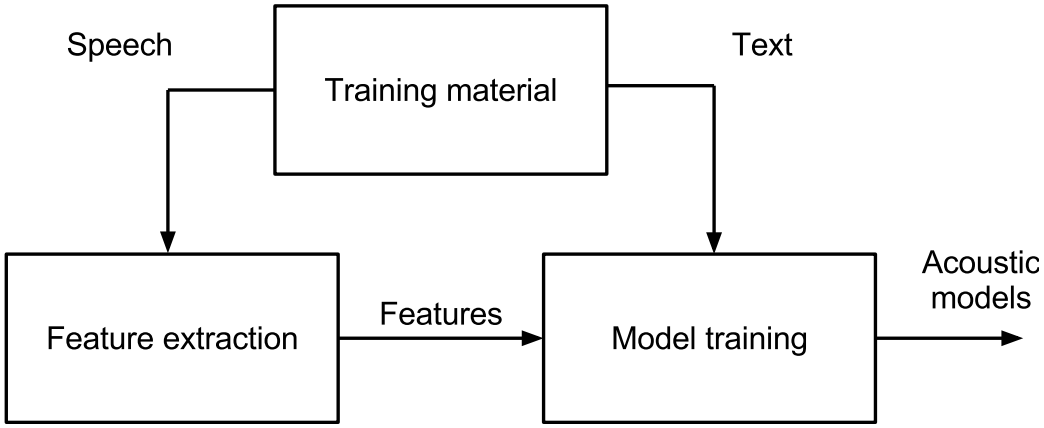


Figure 2.1: Training of acoustic models for an automatic speech recognizer

In Figure 2.2 the testing procedure of an ASR is shown. Again, feature extraction is performed on the speech to obtain the same kind of features as used during training. The acoustic models obtained during training are then used to calculate the probability for each of the models producing the features. The acoustical probabilities are then combined with any grammatical constraints, and in the case of phoneme models; lexical information, and a decision is made about what is said.

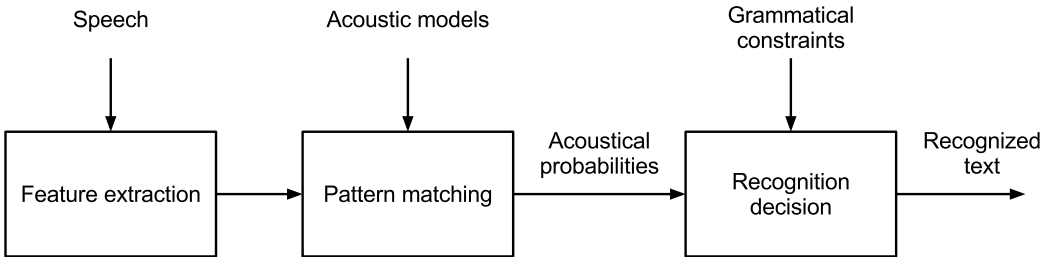


Figure 2.2: Recognition procedure of an automatic speech recognizer

The Hidden Markov Model Toolkit (HTK) was used as the framework to create the ASR system in this thesis. HTK is an open source toolkit for building and manipulating hidden Markov models, primarily made for speech recognition [Young et al., 2006]. It has implemented support for most of the algorithms used, and being open source, makes it easy to implement new algorithms for testing.

Chapter 3

Performance estimates

3.1 Word error rate

The performance of a recognizer is estimated primarily based on the achieved Word error rate (WER). The WER is defined in (3.1), where N_{sub}, N_{ins}, N_{del} and N_{tot} represents the number of substitutions, insertions, deletions and total number of words, respectively.

$$\text{WER} = 100\frac{N_{sub} + N_{ins} + N_{del}}{N_{tot}} \tag{3.1}$$

While the maximal recognition rate that can be achieved is in itself interesting, the robustness of the recognizer is also essential. Because the adaptation is done during training, the recognizer is adapted for a class of speakers, not a specific speaker. This means that no information about the specific speaker is available at the time of adaptation. The experiments in this paper are tested on a limited set of test data, but attempts are made to evaluate how well the recognizer would handle a different set of speakers.

3.2 Confidence intervals

In order to test whether the results achieved are statistically significant, we want to find an interval inside which we can, with some degree of certainty, say that the true error rate can be found. In order to compare two algorithms against the same test set, McNemar's test was used [Gillick and Cox, 1989]. This test was used to estimate the confidence intervals for different test sets. The calculation of the lower and upper limit of the confidence interval is shown in (3.2) and (3.3), respectively [Harborg, 1990].

$$a_1(n_e) = \frac{n_e - 0.5 + 0.5u_{a/2}^2 - u_{a/2}\sqrt{0.25u_{a/2}^2 + \frac{(n_e-0.5)(n-n_e+0.5)}{n}}}{n + u_{a/2}^2} \tag{3.2}$$

$$a_2(n_e) = \frac{n_e + 0.5 + 0.5u_{a/2}^2 + u_{a/2}\sqrt{0.25u_{a/2}^2 + \frac{(n_e+0.5)(n-n_e-0.5)}{n}}}{n + u_{a/2}^2} \tag{3.3}$$

n is the total number of words in the test set, n_e is the total number of errors. $u_a/2$ is defined by $\Pr(\mathcal{N}(0,1) > u_{a/2}) = a/2$, where a is the significance level of the test.

McNemar's test requires that each test is independent. Proper language models change the probability of what words should follow based on the previous words. Because of this dependence, McNemar's test cannot be used to calculate confidence intervals for WER when using proper language models.

Chapter 4

Mel-Frequency Cepstral Coefficients

In order to create a statistical model of different sounds, some form of representation of the speech signal is needed. Ideally, this representation should be constant for the same phoneme across all speakers, different from other phonemes, and easy to extract. These are features created from the speech signal, used to create the acoustical models and used in the pattern matching procedure against the acoustical models in Figure 2.1 and 2.2.

The waveform itself is a bad choice, since variations are very large, even between different pronunciations from the same speaker. Several types of features have been attempted in speech recognition, with different degrees of success. In these experiments, MFCCs are used. These features are widely used in speech recognition and have been shown to perform well for speech recognition tasks [Davis and Mermelstein, 1980]. They are based on the short-time spectral envelope of the signal, are easy to calculate, and provide a simple and effective way of performing frequency warping, which is an integral part of these experiments.

Figure 4.1: Process of creating Mel-Frequency Cepstral Coefficients

The calculation of MFCCs is shown in Figure 4.1. The signal is divided into many short, overlapping, windows, using a Hamming window. The absolute value of the Digital Fourier Transform (DFT) is taken, and these are run through a triangular Mel filterbank. This is a triangular filterbank spaced according to the Mel scale [Stevens et al., 1937]. The logarithm is

then taken, and a Discrete Cosine Transform (DCT) is performed. In HTK (4.1) is used as the DCT [Young et al., 2006].

$$c_i = \sqrt{\frac{2}{N}} \sum_{j=1}^{N} m_j \cos\left(\frac{\pi i}{N}(j - 0.5)\right) \tag{4.1}$$

Chapter 5

Statistics

Once the MFCCs have been extracted from training data, certain statistical methods are used to create the acoustical models from this data. These models are in turn used to recognize new MFCCs. The Hidden Markov Model (HMM) and Gaussian Mixture Model (GMM) are both an integral part of many ASR systems, and are used to create the acoustical models in these experiments.

5.1 Hidden Markov Models

A Markov chain is a model of a random process. It has a finite number of observable outputs, with a state representing each output. The special thing about a Markov chain is that the probability of a certain state relies only on the previous state, thus using a minimal amount of memory without being completely memoryless.

A HMM is essentially a Markov chain, except that the state can only partially be determined by observations. This means that each state does not represent one particular output, but has a certain probability of giving each observable output. The fact that each state has an output probability of different observations means that the actual state sequence from a set of observations is unknown, thus the name *Hidden* Markov Model.

A HMM is defined by: $\mathbf{O}$, the observable output alphabet; $\mathbf{\Omega}$, a set of states; $\mathbf{A}$, a transition probability matrix representing the probability of transitioning from one state to another; $\mathbf{B}$, an output probability matrix representing the probability of each state generating a certain output; and $\boldsymbol{\pi}$, an initial state probability vector representing the probability of starting in each state.

In these experiments, left-to-right HMMs are mainly used. This means

that each state can only jump to one other state in addition to being able to jump to itself, and every state has to be visited. This way, every state is visited sequentially from start to finish. The only exceptions to this are the models for silence, which can jump more freely between states.

Because the observable output alphabet, $\mathbf{O}$, is not a discrete alphabet but a continuous set of observable outputs of MFCCs, the output probability matrix, $\mathbf{B}$, needs to represent a continuous output probability. To represent these output probabilities, GMMs are used.

5.2 Gaussian Mixture Models

GMMs are used to model continuous output probabilities in the HMM. Because of their ability to model almost any distribution function parametrically, GMMs are very effective in this sense.

The output distribution is given in (5.1). Here, $b_j(\mathbf{o})$ is the probability for state j to generate observation $\mathbf{o}$, where $\mathcal{N}(\mathbf{o}; \mu, \mathbf{\Sigma})$ is a multivariate Gaussian distribution as shown in (5.2) [Huang et al., 2001]. $b_j(\mathbf{o})$ are used as the elements of the output probability matrix $\mathbf{B}$ in the HMM.

$$b_j(\mathbf{o}) = \sum_{m=1}^{M} c_{jm} \mathcal{N}(\mathbf{o}; \mu_{jm}, \mathbf{\Sigma}_{jm}) \tag{5.1}$$

$$\mathcal{N}(\mathbf{o}; \mu, \mathbf{\Sigma}) = \frac{1}{\sqrt{(2\pi)^n |\mathbf{\Sigma}|}} e^{-\frac{1}{2}(\mathbf{o}-\mu)^\mathsf{T} \mathbf{\Sigma}^{-1} (\mathbf{o}-\mu)} \tag{5.2}$$

Thus, when the MFCC features are to be tested during recognition to determine which model is more likely to have produced the sound, the GMMs produce a certain probability for each MFCC vector, in each HMM state, representing the probability that that state would produce that MFCC vector.

Chapter 6

Vocal tract length normalization

VTLN is a method used to compensate for spectral differences between different speakers [Lee and Rose, 1996]. As mentioned in Chapter 4, the features used to represent speech rely heavily on the spectral information in the speech signal. This means that spectral differences between speakers could have a large impact on the recognition rate, so adjusting for these differences would be important. VTLN is based on spectral warping, essentially remapping certain frequencies to others.

6.1 Warping methods

The warping method is a way of mapping an original frequency to a new, warped frequency. There are several methods used when performing frequency warping for VTLN. In these experiments the pieciewise linear and bilinear warping methods were used.

The amount of warping performed depends on a warp factor, α. Because of the differences in the warping methods, the range of warp factors are different and not directly comparable. To distinguish between bilinear and piecewise linear warp factors, the bilinear warp factors are denoted as α_b and the piecewise linear warp factors as α_{pw}.

6.1.1 Bilinear warping

The bilinear warping methods make use of the bilinear transformation of filters to digital filters. The bilinear filter is implemented as an Infinite Impulse Response (IIR) filter as described by the z-transform in equation

(6.1).

$$\hat{z} = m(z) = \frac{z^{-1} - \alpha_b}{1 - \alpha_b z^{-1}} \tag{6.1}$$

α_b is the warp factor for the bilinear transform. It is required that $|\alpha_b| < 1$ for the filter to be stable. This gives a frequency transformation as shown in equation (6.2) [Oppenheim and Johnson, 1972].

$$\hat{\Omega} = \arctan\left(\frac{(1 - \alpha_b^2)\sin\Omega}{(1 + \alpha_b^2)\cos\Omega - 2\alpha_b}\right) \tag{6.2}$$

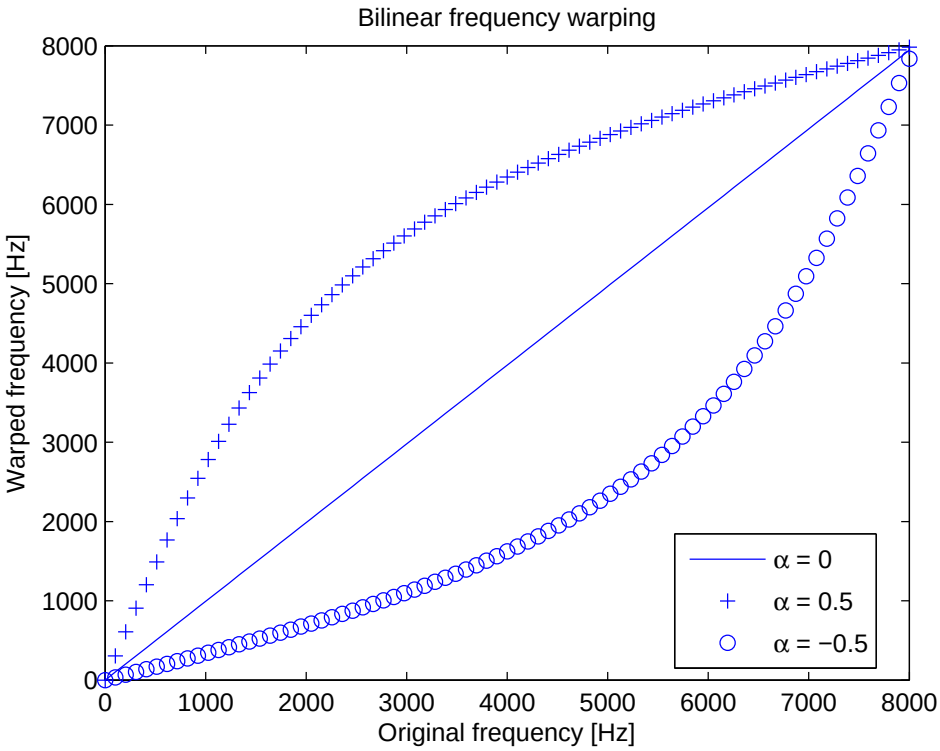


Figure 6.1: Frequency warping using the bilinear method

Figure 6.1 shows the result of bilinear warping with values of α_b above and below zero, compared to the unwarped case where $\alpha_b = 0$. As can be seen from the plot, a value of α_b below zero results in an increased resolution at low frequencies and generally moves the spectrum to lower frequencies, except at the end points. With a warping factor above zero, this is reversed with a higher resolution at high frequencies and moving the spectrum to higher frequencies than the unwarped.

6.1.2 Piecewise linear warping

Piecewise linear warping is a warping method where the spectrum is divided into parts and each part is multiplied by a constant warp factor. The HTK implementation for piecewise linear warping was used. HTK accepts the warp factor as $\frac{1}{\alpha_{pw}}$. Because of this, the piecewise warping factors in the results are always given as $\frac{1}{\alpha_{pw}}$, to avoid large decimals or rounding errors. The HTK implementation of piecewise linear warping needs some explanation, as it does involve a few quirks.

The spectrum is divided into three parts by cutoff frequencies. α_{pw} is the warp factor used in the main area of the spectrum, and the warped frequencies in this area are calculated as shown in (6.3).

$$f_{warped} = \alpha_{pw} f_{original} \tag{6.3}$$

The provided cutoff frequencies, f_{cu} and f_{cl}, are not used directly as cutoff frequencies, but are calculated based on both the provided cutoff frequencies and the warp factor used. The actual cutoff frequencies used, c_u and c_l, are calculated from (6.4) and (6.5). Because of the way the cutoff frequencies are calculated, they may lie outside the used frequency spectrum.

$$c_u = \frac{2 f_{cu}}{1 + \alpha_{pw}} \tag{6.4}$$

$$c_l = \frac{2 f_{cl}}{1 + \alpha_{pw}} \tag{6.5}$$

The calculation of the warp factors used above and below the main area are shown in (6.6) and (6.7), respectively. f_{min} and f_{max} are the minimum and maximum frequencies used, and unless specified these are based on the bandwidth of the signal, in which case f_{min} is zero. (6.7) shows that α_l will then become equal to α_{pw}, so the same warp factor is used below the lower cutoff frequency as in the main area.

$$\alpha_u = \frac{f_{max} - c_u \alpha_{pw}}{f_{max} - c_u} \tag{6.6}$$

$$\alpha_l = \frac{c_l \alpha_{pw} - f_{min}}{c_l - f_{min}} \tag{6.7}$$

From (6.4) it is apparent that even if f_{cu} is below f_{max}, warp factors below a certain threshold will result in a cutoff frequency, c_u, above f_{max}. The result of this is that parts of the upper spectrum is removed, because

the upper cutoff is never reached. When f_{min} is zero, this will result in a completely linear warping.

6.2 Warping implementation

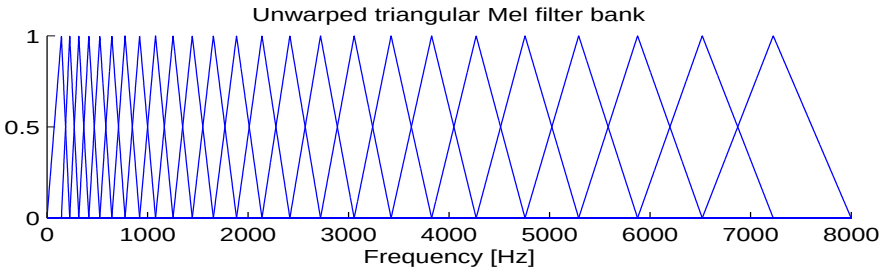
As shown in Figure 4.1, the process of creating MFCCs involve using a triangular Mel filterbank. The frequency warping is implemented by changing the positions of the Mel triangular filters. Figure 6.2 compares the Mel triangular windows on an unwarped scale to warped scales using bilinear and piecewise linear warping.

As Figure 6.1 shows, a warp factor of $\alpha_b < 0$ gives a warped frequency that is lower than the original frequency, except at the endpoints, as well as an increased resolution at low frequencies. Figure 6.2(b) shows this applied to the Mel triangular windows.

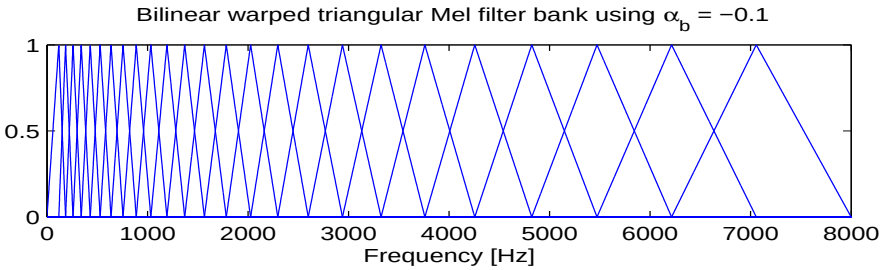
In Figure 6.2(c) the Mel triangular windows when using piecewise warping with a warp factor of $\frac{1}{\alpha_{pw}} = 1.24$ are shown. Here it is apparent that the combination of the warp factor and provided cutoff frequency has resulted in a cutoff frequency above the maximum frequency, resulting in the upper part of the spectrum not being covered by the triangular windows.

Once the warped feature vectors have been calculated, they are treated as if they were placed as the unwarped triangular Mel filters. Because of this, the warp factors are opposite of what would be used if warping was used directly on the signals themselves.

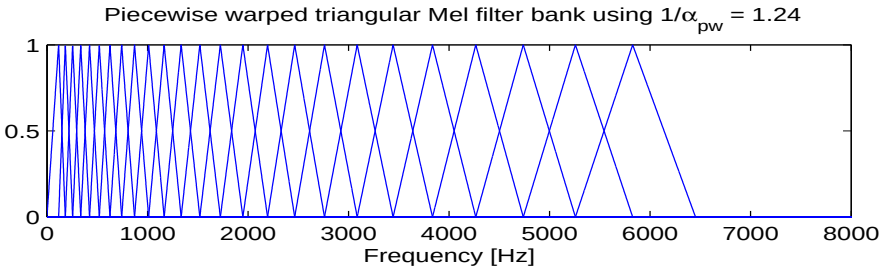
VTLN is most often used during both training and recognition. When done this way, the models are trained iteratively and the warping is decided based on the former models to normalize them with regards to each other [Potamianos and Rose, 1997]. When used during recognition, an attempt is made to estimate the best warp factor for the specific speaker. Several methods exist to estimate the best warp factor [Faria and Gelbart, 2005, Loof et al., 2006], but all of them require a certain amount of extra computation during recognition by both estimating a warp factor and performing the VTLN. Here, the focus is on using VTLN during training for a whole class of people, so the problem of estimating the best warp factor for a specific speaker is not important.



(a) Unwarped Mel triangular windows



(b) Bilinear Mel triangular windows



(c) Piecewise Mel triangular windows

Figure 6.2: Mel triangular windows

Chapter 7

Cepestral mean normalization

CMN is a simple and effective way of increasing the robustness of a ASR. It subtracts the mean of cepstral vectors. This can help to reduce the spectral effects of different microphones and audio channels.

CMN is performed by first finding the mean of a set of cepstral vectors as shown in (7.1), then subtracting this mean from the cepstral vectors to obtain a normalized cepstral vector, $\hat{\mathbf{x}}_t$, instead, as shown in equation (7.2) [Huang et al., 2001].

$$\bar{\mathbf{x}} = \frac{1}{T} \sum_{t=0}^{T-1} \mathbf{x}_t \tag{7.1}$$

$$\hat{\mathbf{x}}_t = \mathbf{x}_t - \bar{\mathbf{x}} \tag{7.2}$$

In HTK the mean is calculated across each input file [Young et al., 2006]. This is likely to make CMN more effective in the offline tests performed than in live recoginition, where the mean would have to be calculated from a more limited set of data.

Chapter 8

Corpus

Several different speech corpora were used in these experiments, each for different purposes. Two large vocabulary corpora, one of adults and one of children, as well as a connected digits corpora of both children and adults, was used.

The two large vocabulary corpora, CMU Kids and TIMIT, were originally used in the specialization project.

8.1 CMU Kids

The Carnegie Mellon University (CMU) Kids corpus is a database of children's speech [Eskenazi et al., 1997]. It contains recordings of 76 children aged 6 to 11, with 24 male and 52 female speakers. There are 356 different sentences used with 878 distinct words.

The corpus was recorded using a Sennheiser headset. Recordings were done on site in a computer room at the school of the children. This resulted in a certain amount of background noise normal in a school environment, such as changing classes. The children read sentences from Weekly Reader, a color reading supplement given out to children. Reading the sentences eliminates any potential imitation one would obtain if the children were asked to repeat sentences from adults; however, it also creates new problems concerning the children's reading abilities.

The corpus came with the utterances divided into two groups, utterances that correctly follow the intended sentences, and utterances containing one or more divergences from the intended sentences. Of a total of 5180 utterances, 4344 contains one or more divergences. This leaves 836 sentences that are properly pronounced. In these experiments, only the properly pronounced recordings were used in testing, in order to limit the influence of the language

model. A word-pair language model was used for this corpus.

Table 8.1: CMU Kids sets

Set name	Speakers	Recordings	Estimated length
CMU Testing set	51	829	48 m
CMU Training set	40	2933	250 m

The sets used from the CMU corpus is shown in Table 8.1. As the number of speakers show, there is some overlap between the speakers in each set. This is due to the testing set using all the properly pronounced sentences, thus encompassing too many of the speakers to avoid any overlap.

8.2 TIMIT

For adult speech the TIMIT Acoustic-Phonetic Continuous Speech Corpus was used [Garofolo et al., 1993]. This is a database of 630 speakers of American English distributed across the seven major dialect regions of the United States of America, as well as one for people who moved around a lot and have no distinct dialect. The TIMIT corpus contains 70% male speakers and 30% female speakers.

The sentences in the TIMIT corpus is divided into three different sets, each for a specific purpose. One set contained sentences designed to expose differences in the various dialects present in the database. Another set was created to maximize the coverage of phoneme pairs. The third set contains sentences from other speech corpora, used to create diversity in sentence types and phonetic contexts. While the sentences used make sense, many are awkward and are unlikely to be used in normal speech.

Table 8.2: TIMIT sets

Set name	Speakers	Recordings	Estimated length
TIMIT Testing set	168	1680	80 m
TIMIT Training set	462	4620	220 m

The corpus is a large vocabulary database. There are 6147 distinct words used in 2342 sentences. The language model used for the TIMIT database is a word-pair model.

8.3 TIDigits

TIDigits is a speech corpus containing recordings of connected digits from both children and adults, male and female. There are in total 326 speakers, and their distribution by age and gender is shown in table 8.3.

Table 8.3: TIDigits age and gender distribution

Category	Number	Age range [years]
Man	111	21-70
Woman	114	17-59
Boy	50	6-14
Girl	51	8-15

Recordings were made in an acoustically treated sound room, using a Electro-Voice RE-16 Dynamic Cardioid microphone. The corpus is originally recorded with a sampling rate of 20 kHz. The recordings were resampled to 16 kHz to keep the results more comparable to the CMU/TIMIT task.

This corpus does not contain any transcriptions apart from information about what numbers are spoken in each recording. Since phoneme level transcriptions with timing details were required in some experiments, forced alignment was used in HTK to create these.

For the TIDigits recognition tasks, a simple word-loop language model was used, where every word has equal probability of following any word. This makes it possible to use McNemar's test to estimate the confidence intervals of the test sets from TIDigits. The 99% and 95% confidence intervals for the TIDigits testing sets are shown in figures A.1 and A.2.

Table 8.4: TIDigits sets

Set name	Speakers	Recordings	Estimated length
TIDigits adult testing set	113	8700	223 m
TIDigits adult training set	112	8623	220 m
TIDigits child testing set	50	3847	116 m
TIDigits child training set	51	3926	117 m

Chapter 9

Experiments and results

9.1 Setup

The general setup was the same across all experiments.

13 MFCCs were used, including an energy coefficient, as well as the delta and acceleration (first and second order derivatives) of these coefficients, for a total of 39 coefficients.

All recordings initially had a sampling rate of 16 kHz. A 10 ms frame rate was used, with a 25 ms window size. Hamming windows were used, as well as a pre-emphasis coefficient of 0.97. 26 filterbank channels were used, N in (4.1). 16 mixtures were used.

When using phoneme models, 3 states are used per phoneme. All experiments included a 3 state silence model and a 1 state model for short pause, bound to the middle state of the silence model.

The upper and lower cut off frequencies used in piecewise linear warping was set to 7500 Hz and 500 Hz, respectively. According to (6.4), this will result in a completely linear warping from a warp factor above $\frac{1}{\alpha_{pw}} = 1.14$.

9.2 Initial problem

The initial results from the specialization project were gained by using the TIMIT corpus to train a recognizer using piecewise linear VTLN, to adapt it to work better for recognizing children's voices in the CMU Kids corpus. The best achieved result was a WER of 80.35, a reduction of 11% from the initial WER of 90.67, using unwarped adult training data.

Because these experiments were carried out using two separate databases, recorded using different equipment and at different locations, the acoustical differences stemming from differences in recording equipment and location

between the recordings are significant. In addition, the CMU kids corpus that was used as an example of children's speech is a large vocabulary corpus with sentences based on children's books. This is problematic in terms of finding a good language model. To reduce these problems in order to focus more on matching the acoustical properties of speech between children and adult speakers, the experiments in this thesis were first performed on the TIDigits corpus. Several new methods were tried to further reduce the errors for children, before the best performing methods were tested against the TIMIT and CMU corpus task, to see if a similar performance gain could be achieved in the large vocabulary task.

9.3 TIDigits

Because of the high WER achieved in the specialization project, the initial recognition attempts using the TIDigits corpus used whole word models. Because this is a connected digits corpus containing only 11 words, this is a feasible task and it is made possible by both the training data and the testing data containing the same words. The whole word models are longer than phoneme models and are likely to be more distinct from each other.

The experiments were carried out with two versions of word models with a different amount of states in the HMM model, one where all words were modelled with 10 states each, and one with a variable amount of states per model to reflect the variable amount of phonemes for each word. The number of states per word used in the variable-state models are shown in Table 9.1.

Table 9.1: Number of states and phonemes per word in the variable-state word models

Word	States	Phonemes
oh	3	1
zero	8	4
one	6	3
two	5	2
three	6	3
four	6	3
five	6	3
six	8	4
seven	10	5
eight	5	2
nine	6	3

Figure 9.1 shows the results of using bilinear warping on the TIDigits adult training set and running recognition on the TIDigits child testing set. The results using piecewise linear warping are shown in Figure 9.2. The detailed results are given in Appendix B.1. In Table 9.2 the results of training on the TIDigits child training set are shown for the different model types.

Table 9.2: WER of a recognizer trained on TIDigits child training set tested against the TIDigits child testing set, using different models

Variable-state	10-state	Phoneme
2.41	0.68	1.71

Of the word models, the 10-state models clearly work best. The variable-state models suffer from a lot of insertion errors because of the `oh` word for zero, which has the least amount of states. The phoneme model does not work as well as the 10-state models, but obtains better results than the variable-state models.

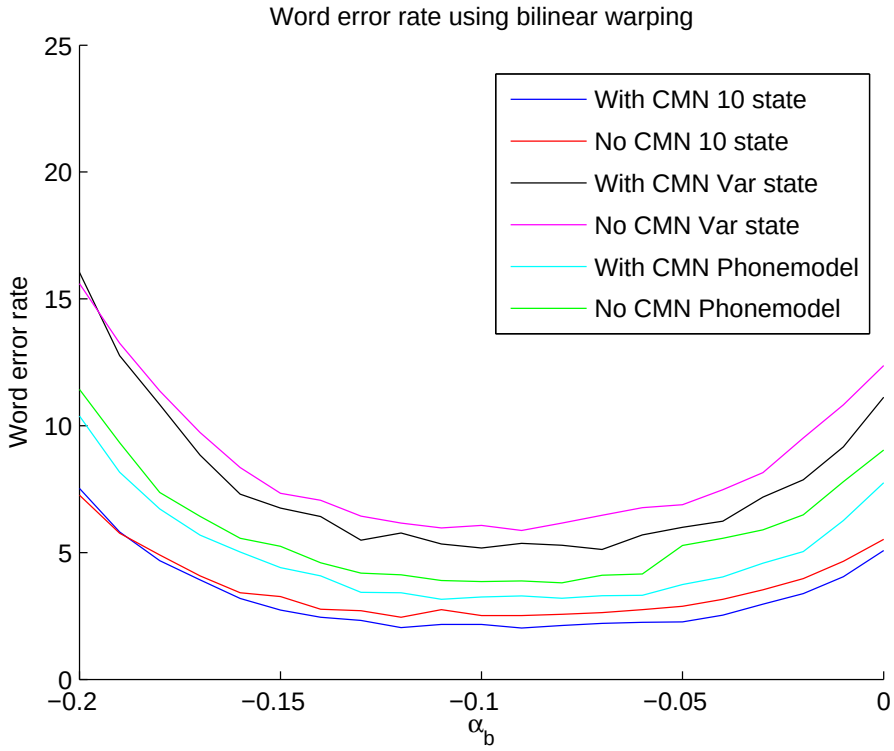


Figure 9.1: Word error rate using bilinear vocal tract length normalisation

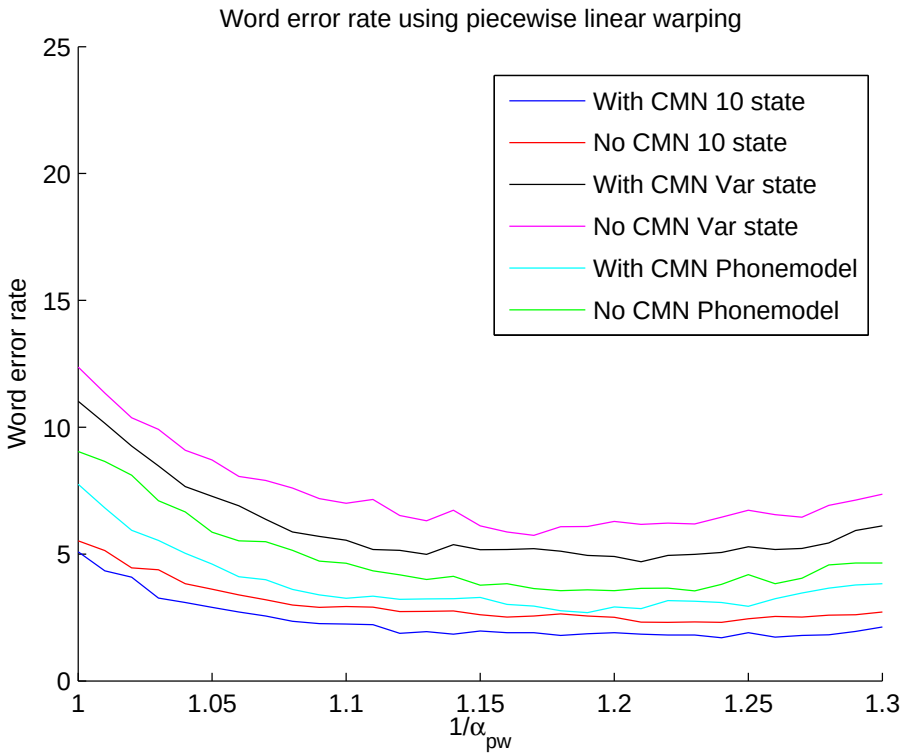


Figure 9.2: Word error rate using piecewise linear vocal tract length normalisation

As the figures show, all the models benefit from the use of CMN, so following experiments all include CMN. The piecewise linear warping method yields better results than the bilinear method. The best achieved WER on the phoneme models was 2.69 for the piecewise linear method and 3.16 for the bilinear method. 3.16 is outside the 99% confidence interval for 2.69 on the TIDigits child testing set, shown in Figure A.1. The range of warp factors where the WER is reduced compared to the unwarped version is also larger when using piecewise linear warping. The WER of 2.69 is a reduction of over 70% compared to the WER of 9.04 when using adult data without CMN and VTLN.

9.3.1 Phonetic class warping

Because the variations in frequencies from different speakers comes in part from the differences in vocal tract length, it would be natural that these differences are smaller for phonemes that are not as influenced by the vocal tract, such as fricatives. It has been suggested that VTLN can cause some

spectral distortion, especially to the silence models [Elenius, 2010]. By not adapting the complete training material with one global warp factor, but using class-dependent warp factors, the silence models can be left unwarped, avoiding any distortions.

To adjust for this, HTK was modified to allow for phoneme-specific warp factors. Because of the huge search space in finding the best warp factor for each phoneme when using phoneme-specific warp factors, finding the best warp factor for each phoneme would be a very difficult task. Instead, the phonemes were divided into three classes depending on the significance of the vocal tract when creating the sound. One class with a low warp factor, one with a medium warp factor, and one with a high warp factor. The best warp factor was found using an exhaustive search with with warp factors, $\frac{1}{\alpha_{pw}}$, between 1.00 to 1.30 in steps of 0.02 for all classes, with the assumption that the highest class, containing vowels, should have a higher warp factor than the medium class, containing nasals and glides, which should have a higher warp factor than the lowest class, containing fricatives.

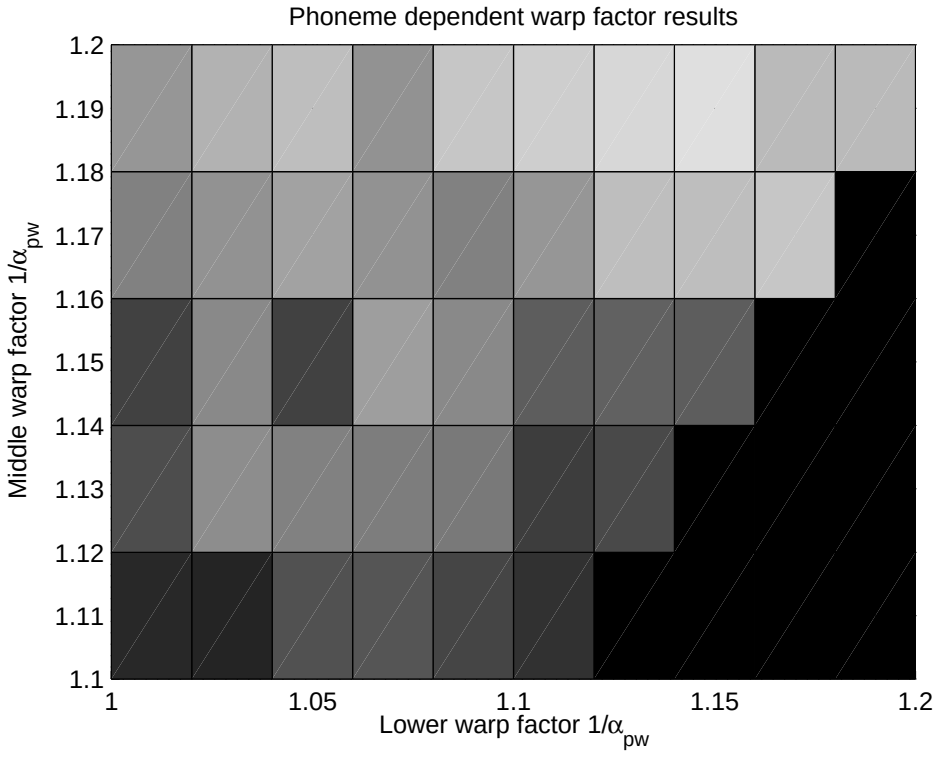


Figure 9.3: Results of phonetic class warping when holding the highest warp factor constant at $\frac{1}{\alpha_{pw}} = 1.2$. Brighter colors represents a lower WER

Due to the problem of representing four dimensions on paper, Figure 9.3 shows the results when having the highest warped class constant at $\frac{1}{\alpha_{pw}} = 1.2$ and varying the warp factor of the two other classes. The brighter the color, the lower the WER.

As the figure shows, the WER is generally lower closer to the top right, where the warp factors are closer together. This general form holds true when changing the higher warp factor as well. The best result gained is a WER of 2.64, which is barely lower than using a constant warp factor and well inside the 95% confidence interval.

9.3.2 Multiple warp factors

As the constant warp factor results show, there is a certain amount of ripple in the results. This indicates a lack of robustness, where the recognition rate of each speaker relies heavily on the warp factor chosen during training. In order to increase the robustness of the recognizer, the TIDigits adult training set was warped with several different warp factors and then combined to train the recognizer, creating a multi set.

Table 9.3: WER of a recognizer trained on the TIDigits adult training set with several warp factors, using CMN and phoneme models. Tested against TIDigits child testing set and TIDigits adult testing set

Warping method	Range	Step	Sets	WER child	WER adult
Piecewise linear	1.00 - 1.30	0.03	11	2.71	2.36
Piecewise linear	1.00 - 1.30	0.05	7	2.90	2.26
Piecewise linear	1.10 - 1.30	0.03	7	2.60	3.75
Piecewise linear	1.10 - 1.30	0.05	5	2.56	3.57
Bilinear	1.00 - 1.20	0.03	7	2.76	2.39
Bilinear	1.00 - 1.20	0.05	5	3.01	2.42
Bilinear	1.05 - 1.15	0.03	4	2.89	3.62
Bilinear	1.05 - 1.15	0.05	3	2.96	3.66

Table 9.3 shows the result of using the same training material, with different combinations warp factors, using phoneme models. While these sets do not provide a large decrease in WER towards children's speech compared to the single warp factor recognizers, they do retain their recognition rate much better than the single warp factor recognizers, as shown in table B.7.

9.3.3 Recognizer comparison

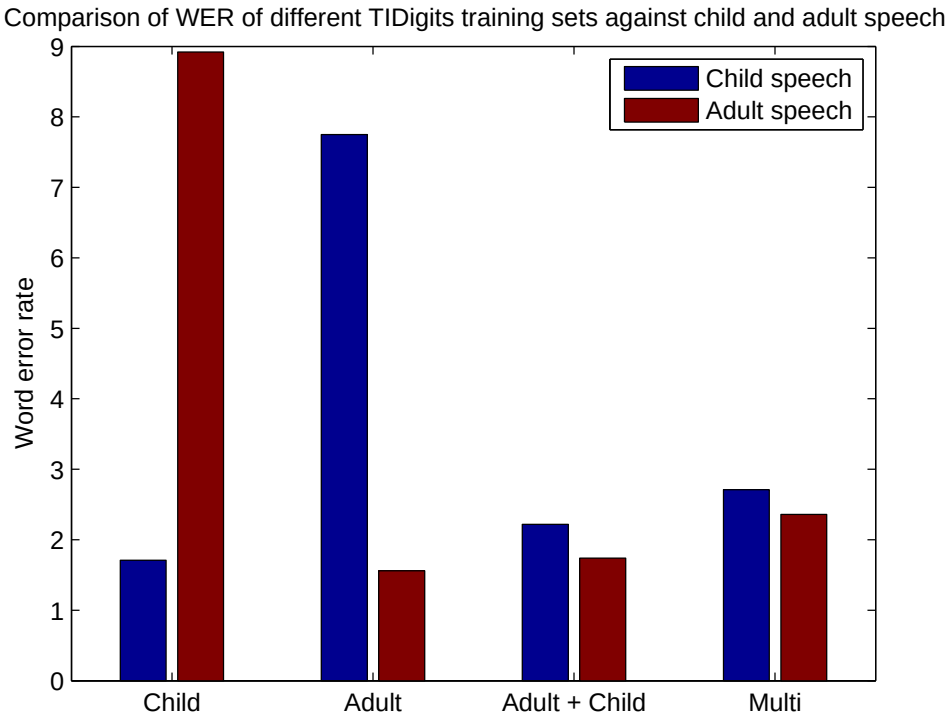


Figure 9.4: Comparison of word error rates of different TIDigits training sets against both child and adult speech. The different sets used are: Child, the TIDigits child training set; Adult, the TIDigits adult training set; Adult+Child, the TIDigits adult and child training sets; and multi, the set using 11 warp factors between 1.00 and 1.30 with piecewise linear warping made from the TIDigits adult training set.

Figure 9.4 gives a comparison of the different sets being tested against both adult and child speech. The recognizers trained on only adult or only child material perform best on their respective testing sets, but definitely perform worst on the other set. The recognizer trained on combined adult and child data perform fairly well on both sets, with an 12% increase in WER against adult speech, compared to the recognizer trained on adult material, and a 30% increase against child speech, compared to the recognizer trained on child material.

The multi training set performs worse than the combined set, but otherwise has much the same form, with a slightly higher WER on child speech

than adult speech. The increase in WER compared to the combined set is 22% and 36% for child and adult speech, respectively.

9.4 TIMIT and CMU

The improvements made to the recognition rate on the TIDigits database was transferred back to the recognizer trained on TIMIT. Table B.8 shows the results of applying both VTLN and CMN. This shows a best WER of 46.70, which a 41.9% reduction from the best result of the specialization project, without CMN, and a 48.5% reduction from the initial unwarped WER without CMN.

Table 9.4: Results of using multiple warp factors in one recognizer with adult data from TIMIT using piecewise linear warping.

Warp factor range	Step	Number of sets	WER TIMIT Test	WER CMU Test
1.00-1.30	0.05	7	22.39	48.95
1.10-1.40	0.05	7	45.17	48.27
0.85-1.30	0.05	10	18.87	55.62
0.85-1.40	0.05	12	19.95	52.86

Table 9.4 shows the result of using the TIMIT training set several times with multiple warp factors, and tested against both adult and child data. These results show a fairly clear tradeoff between the range of warp factors and the recognition rate against each set. One thing to note is that the set with factors between 0.85 and 1.30 shows a 9% increase in the WER compared to the results using an unwarped recognizer, while giving a 32% reduction in WER for the child speech.

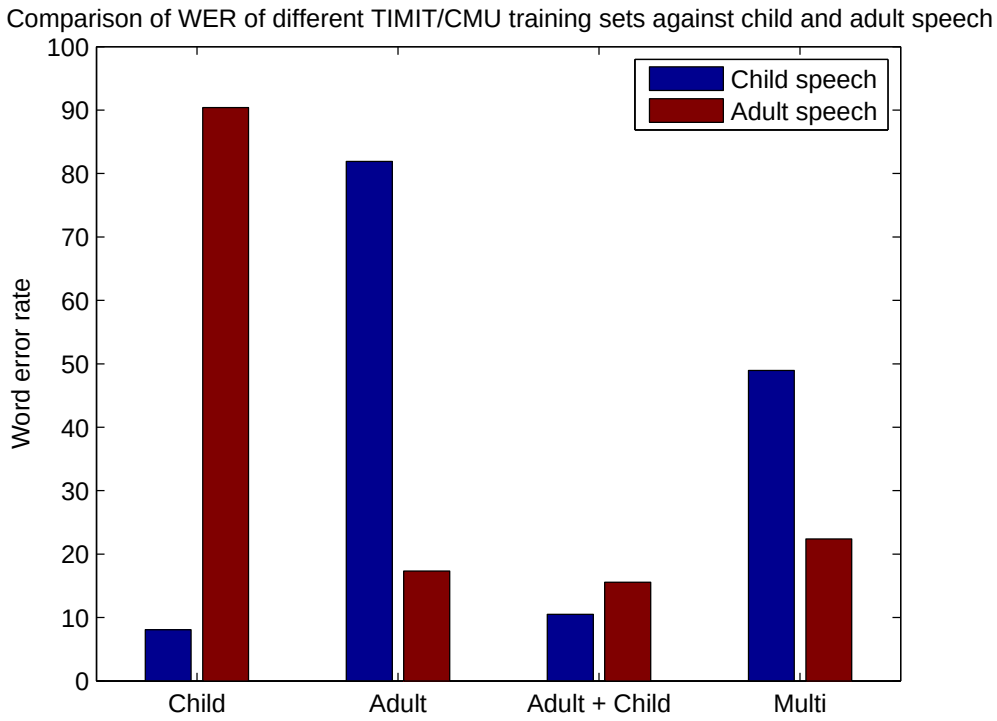


Figure 9.5: Comparison of word error rates of different CMU and TIMIT training sets against both child and adult speech. The different sets used are: Child, the CMU training set; Adult, the TIMIT training set; Adult+Child, the TIMIT and CMU training sets; and multi, the set using 7 warp factors between 1.00 and 1.30 with piecewise linear warping made from the TIMIT training set.

A comparison of different sets from the TIMIT and CMU corpora are shown in Figure 9.5. Training sets from both corpora perform very bad on the testing set from the other corpus when unaltered. Again the combined set performs better than the multi set, and even outperforms the adult set against adult speech. The multi set performs slightly worse than the combined set against the adult data, but performs significantly worse against the child speech, also when compared to the connected digits task.

Chapter 10

Discussion

10.1 Warping methods

The results obtained in this study show that the piecewise linear warping method provided by HTK outperformed the bilinear warping method. As mentioned, because of the way piecewise linear VTLN is implemented in HTK, some warp factors will result in a completely linear warping, removing upper parts of the spectrum.

From Table B.6, we see there is very little difference between $\frac{1}{\alpha_{pw}} = 1.14$ and $\frac{1}{\alpha_{pw}} = 1.15$, where the jump between piecewise linear and linear is made. This indicates that the frequencies above $7000\,Hz$ have almost no effect. Because of this lack of influence of the upper part of the spectrum, it is unlikely that this is the cause of the lower WER from piecewise linear warping that bilinear warping.

10.2 Phonetic class warping

The use of separate warp factors for different phonetic classes only gave a marginal reduction in WER compared to the use of a global warp factor, and performed marginally worse than the best multi set. Although this in line with previous results of using phoneme-specific warp factors [Elenius, 2010, Potamianos and Narayanan, 2003], in [Potamianos and Narayanan, 2003] it was suggested that this might be due to the trouble of estimating good warp factors in such a large search space. Here, the search space was reduced by using separate warp factors for phonetic classes instead of each phoneme, and an exhaustive search was performed, but no significant performance increase could be found.

That the best recognition rates are gained by all the factors being close to each other may indicate that the resolution of the timing details in the phoneme transcriptions is not high enough. If the timing is not detailed enough, the end points of phonemes would be affected by the warp factors of other phonemes. If the accuracy is too low, too much of the phonemes will be affected by other classes and each phoneme class will contain a mixture of several phoneme class warp factors. This could be solved by performing the adaptation on the models instead of the features used for training.

10.3 Robustness

For every test of recognition rate for children, applying VTLN did reduce the WER for a significant area of warp factors. However, within this area, there was always a certain amount of ripple. This indicates, unsurprisingly, that there is variability in child speech.

By training a system on adult speech warped with several different warp factors, the recognition rate towards child speech has been maintained, but the recognition rate for adult speech has been increased significantly, compared to the single warp factor recognizers. This seems to indicate an increased robustness for the system, and it is likely that these systems will retain their recognition rates better for a larger amount of children as well.

10.4 Large vocabulary recognition

The large vocabulary recognition still suffers from a very high WER. While a reduction of 48.5% is significant, the WER of 46.70 is still far too high to be useful in an ASR. The very high increase gained from applying CMN and the large discrepancy between the results of the combined training set compared to the multi training set compared to the connected digits task indicates a large difference between the databases.

10.5 Computational cost

Performing adaptation during training reduces the computational costs of adaptation during testing. It was suggested that using a multi set would reduce the computational time to a fraction of normal VTLN [Elenius, 2010], however, in the large vocabulary task the recognition time of the best multi set compared to the best single warp factor when recognizing children was only reduced by a marginal amount of 6%.

Chapter 11

Conclusion

It has been shown that a combination of VTLN and CMN can significantly reduce the WER for children on systems trained on adult speech, with a reduction of over 70% for a connected-digits task, and almost 50% for a large vocabulary task. By using the same training data, but with multiple warp factors, WER was slightly reduced, but more significantly, it retained a much lower WER for adult speech than using only one warp factor. While this is unlikely to be sufficient to create a recognizer that is satisfactory for both children and adults, it does show that the system is less speaker specific. The better recognition rate for adults indicates a that the system is more robust, and it is likely it will retain a better recognition rate for more children.

This increase in recognition rate and robustness comes as a very low cost. The same training data is used several times, so no extra training data needs to be gathered. The multiplication of the existing training data, CMN and VTLN does increase the computational power needed during training, but power during training is very cheap and it only has to be done once. CMN does require some extra computational power during recognition, but not a significant amount.

It was also shown that a linear frequency warping is a better approximation of the differences between adult and child speech than bilinear warping. It is also likely that, when using a sampling rate of 16 kHz, the frequencies above 7 kHz in the adult speech has little effect on the recognition of child speech.

Bibliography

[Das et al., 1998] Das, S., Nix, D., and Picheny, M. (1998). Improvements in children's speech recognition performance. In *Acoustics, Speech and Signal Processing, 1998. Proceedings of the 1998 IEEE International Conference on*, volume 1, pages 433–436.

[Davis and Mermelstein, 1980] Davis, S. B. and Mermelstein, P. (1980). Comparison of parametric representations for monosyllabic word recognition in continuously spoken sentences. *Acoustics, Speech and Signal Processing, IEEE Transactions on*, 28(4):357–366.

[Elenius, 2010] Elenius, D. (2010). *Accounting for Individual Speaker Properties in Automatic Speech Recognition*. Licentiate thesis, Kungliga Tekniska högskolan.

[Eskenazi et al., 1997] Eskenazi, M., Mostow, J., and Graff, D. (1997). The CMU Kids Corpus.

[Faria and Gelbart, 2005] Faria, A. and Gelbart, D. (2005). Efficient Pitch-Based Estimation of VTLN Warp Factors. *Proc. INTERSPEECH*, pages 213–216.

[Fitch and Giedd, 1999] Fitch, W. T. and Giedd, J. (1999). Morphology and development of the human vocal tract. *J. Acoust. Soc. Am.*, 106(3):1511–1522.

[Fjær, 2010] Fjær, B. G. (2010). *Speech adoption for children using vocal tract length normalization*. Specialization project, Norwegian University of Science and Technology.

[Garofolo et al., 1993] Garofolo, J. S., Lamel, L. F., Fisher, W. M., Fiscus, J. G., Pallett, D. S., Dahlgren, N. L., and Zue, V. (1993). TIMIT Acoustic-Phonetic Continuos Speech Corpus.

[Gillick and Cox, 1989] Gillick, L. and Cox, S. (1989). Some statistical issues in the comparison of speech recognition algorithms. In *Acoustics, Speech and Signal Processing, 1989. ICASSP-89., 1989 International Conference on*, volume 1, pages 532–535.

[Harborg, 1990] Harborg, E. (1990). *Hidden Markov Models applied to Automatic Speech Recognition*. PhD thesis, Norges Tekniske Høgskole.

[Huang et al., 2001] Huang, X., Acero, A., and Hon, H.-W. (2001). *Spoken language processing*. Prentice Hall PTR.

[Kazemzadeh et al., 2005] Kazemzadeh, A., You, H., Iseli, M., Jones, B., Cui, X., Heritage, M., Price, P., Anderson, E., Narayanan, S., and Alwan, A. (2005). TBALL data collection: The making of a children's speech corpus. *Proc. of INTERSPEECH/EUROSPEECH*, pages 1581–1584.

[Lee and Rose, 1996] Lee, L. and Rose, R. C. (1996). Speaker normalization using efficient frequency warping procedures. *Acoustics, Speech, and Signal Processing, IEEE International Conference on*, 1:353–356.

[Lee et al., 1999] Lee, S., Potamianos, A., and Narayanan, S. (1999). Acoustics of children's speech. *J. Acoust. Soc. Am.*, 105(3).

[Loof et al., 2006] Loof, J., Ney, H., and Umesh, S. (2006). Vtln Warping Factor Estimation Using Accumulation of Sufficient Statistics. *Proc. ICASSP-06*, 1.

[Narayanan and Potamianos, 2002] Narayanan, S. and Potamianos, A. (2002). Creating conversational interfaces for children. *Speech and Audio Processing, IEEE Transactions on*, 10(2):65–78.

[Oppenheim and Johnson, 1972] Oppenheim, A. V. and Johnson, D. D. (1972). Discrete Representation of Signals. *Proceedings of the IEEE*, 60(6):681–691.

[Potamianos and Narayanan, 2003] Potamianos, A. and Narayanan, S. (2003). Robust Recognition of Children's Speech. *Speech and Audio Processing, IEEE Transactions on*, 11(6):603–616.

[Potamianos and Rose, 1997] Potamianos, A. and Rose, R. C. (1997). On combining frequency warping and spectral shaping in HMM based speech recognition. In *Acoustics, Speech and Signal Processing, 1997. ICASSP-97., 1997 IEEE International Conference on*, volume 2, pages 1275–1278.

[Shobaki et al., 2000] Shobaki, K., Hosom, J.-P., and Cole, R. A. (2000). The OGI Kids' Speech Corpus and recognizers. *In Proc. of ICSLP*, pages 564–567.

[Stemmer et al., 2003] Stemmer, G., Hacker, C., Steidl, S., and Nöth, E. (2003). Acoustic Normalization of Children's Speech. *Proceedings of European Conference on Speech Communication and Technology*, 2:1313–1316.

[Stevens et al., 1937] Stevens, S. S., Volkmann, J., and Newman, E. B. (1937). A Scale for the Measurement of the Psychological Magnitude Pitch. *J. Acoust. Soc. Am.*, 8(3):185–190.

[Wilpon and Jacobsen, 1996] Wilpon, J. G. and Jacobsen, C. N. (1996). A study of speech recognition for children and the elderly. *Acoustics, Speech, and Signal Processing, IEEE International Conference on*, 1:349–352.

[Young et al., 2006] Young, S., Evermann, G., Gales, M., Hain, T., Kershaw, D., Liu, X. A., Moore, G., Odell, J., Ollason, D., Povey, D., Valtchev, V., and Woodland, P. (2006). The HTK Book. `http://htk.eng.cam.ac.uk/docs/docs.shtml`.

Appendix A

Confidence intervals

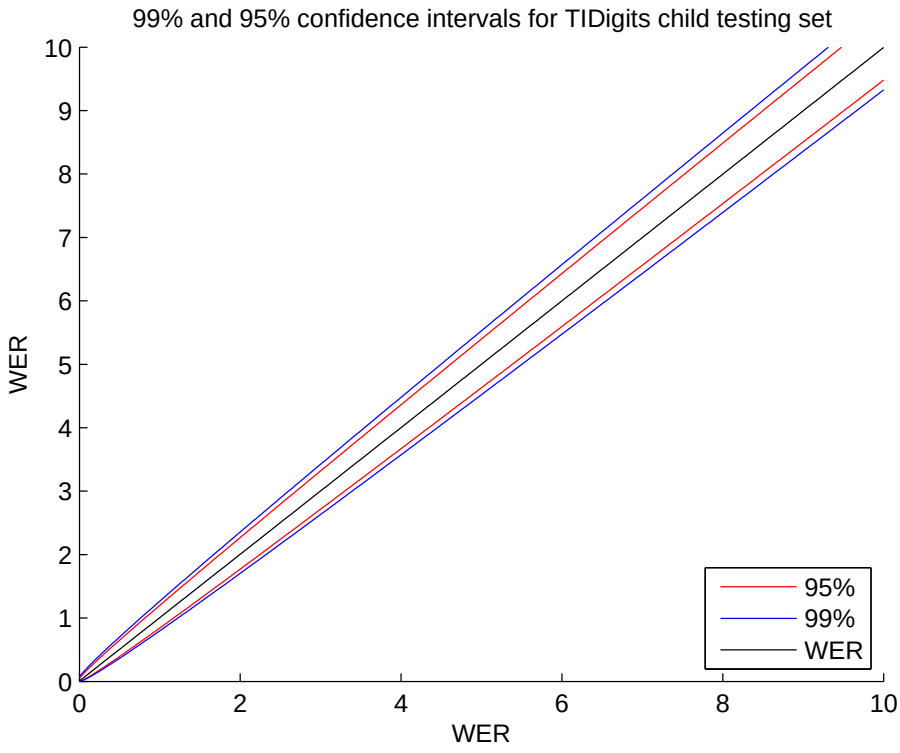


Figure A.1: 99% and 95% confidence intervals for TIDigits child testing set

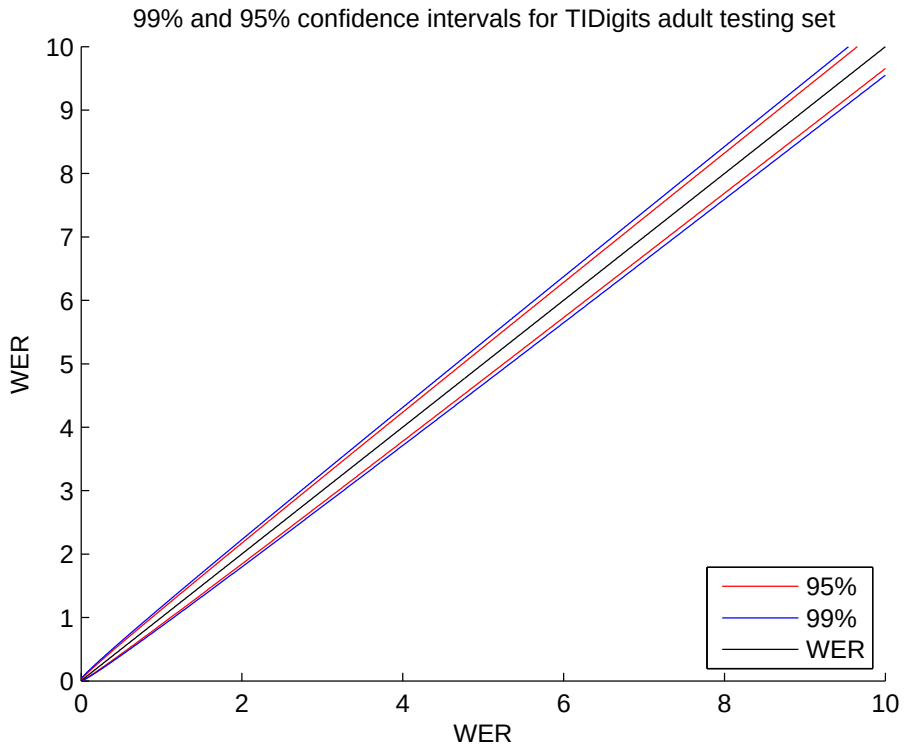


Figure A.2: 99% and 95% confidence intervals for TIDigits adult testing set

Appendix B

Results

B.1 TIDigits

B.1.1 Variable-state models

Table B.1: WER using piecewise linear VTLN with variable-state models, with and without CMN. Trained on the TIDigits adult training set and tested against the TIDigits child testing set

Warping factor ($\frac{1}{\alpha_{pw}}$)	WER with CMN	WER without CMN
1.00	11.02	12.37
1.01	10.16	11.35
1.02	9.26	10.37
1.03	8.48	9.92
1.04	7.66	9.09
1.05	7.28	8.71
1.06	6.90	8.06
1.07	6.37	7.90
1.08	5.87	7.60
1.09	5.70	7.19
1.10	5.55	7.00
1.11	5.18	7.15
1.12	5.15	6.52
1.13	4.99	6.31
1.14	5.37	6.73
1.15	5.17	6.11
1.16	5.18	5.87
1.17	5.21	5.74
1.18	5.11	6.08
1.19	4.95	6.09
1.20	4.91	6.29
1.21	4.70	6.17
1.22	4.95	6.22
1.23	4.99	6.19
1.24	5.06	6.45
1.25	5.29	6.73
1.26	5.18	6.55
1.27	5.22	6.45
1.28	5.44	6.92
1.29	5.93	7.13
1.30	6.11	7.36

Table B.2: WER using bilinear VTLN with variable-state models, with and without CMN. Trained on the TIDigits adult training set and tested against the TIDigits child testing set

Warping factor (α_b)	WER with CMN	WER without CMN
0.00	11.02	12.37
-0.01	9.17	10.82
-0.02	7.87	9.52
-0.03	7.19	8.15
-0.04	6.24	7.48
-0.05	6.00	6.89
-0.06	5.70	6.77
-0.07	5.12	6.47
-0.08	5.29	6.16
-0.09	5.36	5.87
-0.10	5.18	6.07
-0.11	5.34	5.97
-0.12	5.77	6.16
-0.13	5.49	6.44
-0.14	6.42	7.06
-0.15	6.75	7.34
-0.16	7.30	8.35
-0.17	8.84	9.73
-0.18	10.83	11.37
-0.19	12.76	13.25
-0.20	16.04	15.60

B.1.2 10-state models

Table B.3: WER using piecewise linear VTLN with 10-state models, with and without CMN. Trained on the TIDigits adult training set and tested against the TIDigits child testing set

Warping factor ($\frac{1}{\alpha_{pw}}$)	WER with CMN	WER without CMN
1.00	5.09	5.52
1.01	4.34	5.14
1.02	4.09	4.46
1.03	3.27	4.38
1.04	3.09	3.83
1.05	2.90	3.62
1.06	2.72	3.39
1.07	2.56	3.20
1.08	2.35	2.99
1.09	2.26	2.90
1.10	2.24	2.93
1.11	2.22	2.91
1.12	1.88	2.73
1.13	1.94	2.74
1.14	1.84	2.76
1.15	1.97	2.61
1.16	1.90	2.52
1.17	1.90	2.56
1.18	1.79	2.64
1.19	1.86	2.56
1.20	1.90	2.51
1.21	1.84	2.32
1.22	1.81	2.31
1.23	1.81	2.33
1.24	1.70	2.31
1.25	1.90	2.45
1.26	1.73	2.54
1.27	1.79	2.52
1.28	1.82	2.59
1.29	1.95	2.61
1.30	2.13	2.72

Table B.4: WER using bilinear VTLN with 10-state models, with and without CMN. Trained on the TIDigits adult training set and tested against the TIDigits child testing set

Warping factor (α_b)	WER with CMN	WER without CMN
0.00	5.08	5.52
-0.01	4.05	4.66
-0.02	3.38	3.97
-0.03	2.97	3.53
-0.04	2.53	3.16
-0.05	2.27	2.88
-0.06	2.25	2.75
-0.07	2.21	2.63
-0.08	2.13	2.57
-0.09	2.03	2.52
-0.10	2.17	2.52
-0.11	2.17	2.75
-0.12	2.04	2.45
-0.13	2.33	2.71
-0.14	2.45	2.77
-0.15	2.73	3.27
-0.16	3.19	3.42
-0.17	3.92	4.08
-0.18	4.68	4.91
-0.19	5.81	5.76
-0.20	7.53	7.26

B.1.3 Phone model

Table B.5: WER using bilinear VTLN with phoneme models, with and without CMN. Trained on the TIDigits adult training set and tested against the TIDigits child testing set

Warping factor (α_b)	WER with CMN	WER without CMN
0.00	7.75	9.04
-0.01	6.26	7.79
-0.02	5.04	6.49
-0.03	4.58	5.90
-0.04	4.04	5.56
-0.05	3.74	5.28
-0.06	3.32	4.16
-0.07	3.30	4.11
-0.08	3.20	3.81
-0.09	3.29	3.88
-0.10	3.25	3.86
-0.11	3.16	3.90
-0.12	3.42	4.12
-0.13	3.43	4.19
-0.14	4.08	4.60
-0.15	4.41	5.25
-0.16	5.02	5.56
-0.17	5.69	6.43
-0.18	6.72	7.37
-0.19	8.17	9.33
-0.20	10.38	11.43

Table B.6: WER using piecewise linear VTLN with phoneme models, with and without CMN. Trained on the TIDigits adult training set and tested against the TIDigits child testing set

Warping Factor ($\frac{1}{\alpha_{pw}}$)	WER with CMN	WER without CMN
1.00	7.75	9.04
1.01	6.82	8.65
1.02	5.94	8.11
1.03	5.54	7.10
1.04	5.03	6.66
1.05	4.61	5.86
1.06	4.11	5.52
1.07	3.99	5.49
1.08	3.61	5.15
1.09	3.39	4.72
1.10	3.26	4.64
1.11	3.34	4.34
1.12	3.22	4.18
1.13	3.23	4.00
1.14	3.24	4.12
1.15	3.29	3.77
1.16	3.02	3.83
1.17	2.95	3.64
1.18	2.77	3.56
1.19	2.69	3.59
1.20	2.92	3.56
1.21	2.85	3.65
1.22	3.17	3.66
1.23	3.14	3.55
1.24	3.09	3.81
1.25	2.94	4.19
1.26	3.24	3.83
1.27	3.47	4.05
1.28	3.66	4.57
1.29	3.78	4.65
1.30	3.83	4.65

Table B.7: WER using piecewise linear VTLN with phoneme models Trained on the TIDigits adult training set and tested against the TIDigits adult testing set

Warping Factor ($\frac{1}{\alpha_{pw}}$)	WER
1.00	1.56
1.02	1.60
1.04	1.52
1.06	1.69
1.08	2.04
1.10	2.26
1.12	2.67
1.14	3.20
1.16	3.61
1.18	4.67
1.20	5.73
1.22	7.19
1.24	9.59
1.26	12.78
1.28	15.90
1.30	19.68

B.2 CMU

Table B.8: WER against CMU Testing set using piecewise linear VTLN and CMN on a recognizer trained on TIMIT Training set

Training set	WER
1.00	81.91
1.02	75.30
1.04	70.13
1.06	65.02
1.08	59.62
1.10	58.24
1.12	52.72
1.14	53.49
1.16	50.56
1.18	47.69
1.20	47.73
1.22	46.98
1.24	46.70
1.26	47.55
1.28	48.58
1.30	46.98
1.32	48.63
1.34	51.27